

# Protein secondary structure analysis with a coarse-grained model

Gerald R. Kneller<sup>1,2,3\*</sup> and Konrad Hinsén<sup>1,2</sup>

<sup>1</sup>*Centre de Biophys. Moléculaire, CNRS; Rue Charles Sadron, 45071 Orléans, France*

<sup>2</sup>*Synchrotron Soleil; L'Orme de Merisiers, 91192 Gif-sur-Yvette, France and*

<sup>3</sup>*Université d'Orléans; Chateau de la Source-Av. du Parc Floral, 45067 Orléans, France*

The paper presents a geometrical model for protein secondary structure analysis which uses only the positions of the  $C_\alpha$ -atoms. We construct a space curve connecting these positions by piecewise polynomial interpolation and describe the folding of the protein backbone by a succession of screw motions linking the Frenet frames at consecutive  $C_\alpha$ -positions. Using the ASTRAL subset of the SCOPe data base of protein structures, we derive thresholds for the screw parameters of secondary structure elements and demonstrate that the latter can be reliably assigned on the basis of a  $C_\alpha$ -model. For this purpose we perform a comparative study with the widely used DSSP (Define Secondary Structure of Proteins) algorithm.

PACS numbers: 87.15.-v, 87.15.B-, 87.15.bd

Keywords: Protein secondary structure, coarse-grained protein model

## I. INTRODUCTION

Protein secondary structure elements (PSSE) are the basic building blocks of proteins and their form and arrangement is of fundamental importance for protein folding and function. They have been first predicted by Pauling and Corey on the basis of hydrogen bonding [1, 2] and were later confirmed by X-ray diffraction experiments. The localization of PSSEs in protein structure databases is one of the most basic tasks in bioinformatics and various methods have been developed for this purpose. We mention here DSSP (Define Secondary Structure of Proteins)[3] and STRIDE (STRuctural IDentification) [4], which assign PSSEs on the basis of geometrical, energetic and statistical criteria and which are the most widely used approaches. The result are contiguous domains along the amino acid sequence of the protein, which are labeled as “ $\alpha$ -helix”, “ $\beta$ -strand”, etc. There is no precise and universally accepted definition for PSSEs, and therefore each method produces slightly different results. The geometrical variability of these PSSEs, which depends on the global protein fold, is not explicitly considered by these approaches. The more recently published ScrewFit method [5, 6] allows for both assignment and geometrical description of PSSEs. It describes the geometry of the protein backbone by a succession of screw motions linking successive  $C - O - N$  groups in the peptide bonds, from which PSSEs can be assigned on the basis of statistically established thresholds for the local helix parameters. The latter have been derived by screening the ASTRAL database [7], which provides representative protein structure sets containing essentially one secondary structure motif. The ScrewFit description is intuitive

and bears some resemblances with the P-Curve approach proposed by Sklenar, Etchebest and Lavery [8], in the sense that both methods lead to a sequence of local helix axes, the ensemble of which defines an overall axis of the protein under consideration. ScrewFit uses, however, a minimal set of parameters and was originally developed to pinpoint changes in protein structure due to external stress.

The experimental basis for the automated assignment of PSSEs in proteins is X-ray crystallography, which yields information about the positions of the heavy atoms in a protein. Although the number of resolved protein structures increased almost exponentially during the last two decades, the fraction of proteins for which the atomic structure is known is still very small. Low resolution techniques, like electron microscopy, are an additional source of information [9, 10] and in this context the description of PSSEs must be correspondingly simplified, in order to be useful in structure refinement. A natural and commonly used coarse-grained description of proteins is the  $C_\alpha$ -model, where each residue is represented by its respective  $C_\alpha$ -atom on the protein backbone [11]. To our knowledge, Levitt *et al.* were the first to publish a method of secondary structure assignment on the basis of the  $C_\alpha$ -positions[12], and different approaches for that purpose have been published since then [13–15]. Like DSSP and STRIDE, these methods aim at assigning PSSEs on a true/false basis and the underlying models for this decision are not exploited or not exploitable for a more detailed description of protein folds. The motivation of this paper was to develop an extension of the ScrewFit method which works only with the  $C_\alpha$ -positions, maintaining the capability to describe the global fold of a protein by a minimalistic model and to assign PSSEs. The method is described in Section II and two applications are presented and discussed in Section III. A short résumé with an outlook concludes the paper.

\*Electronic address: gerald.kneller@cnrs-orleans.fr

## II. A COARSE-GRAINED MODEL FOR THE FOLD OF A PROTEIN

### A. $C_\alpha$ space curve and Frenet frames

We consider the ensemble of the  $C_\alpha$ -positions,  $\{\mathbf{R}_1, \dots, \mathbf{R}_N\}$ , as a discrete representation of a space curve,  $\mathbf{r}(\lambda) = \sum_{k=1}^3 r_k(\lambda) \mathbf{e}^{(k)}$ , where  $\lambda \in [\lambda_a, \lambda_b]$  and  $\mathbf{e}^{(k)}$  ( $k = x, y, z$ ) are the basis vectors of a space-fixed Euclidean coordinate system. Imposing that

$$\mathbf{r}(\lambda_j) = \mathbf{R}_j, \quad j = 1 \dots N, \quad (1)$$

at equidistantly sampled values of  $\lambda$ ,

$$\lambda_j = \lambda_a + (j-1)\Delta\lambda, \quad \Delta\lambda = (\lambda_b - \lambda_a)/N, \quad (2)$$

we define a continuous space curve by a piecewise polynomial interpolation of the  $C_\alpha$ -positions. The values for  $\lambda_a$  and  $\lambda_b$  are arbitrary and one may in particular choose  $\lambda_a = 0$  and  $\lambda_b = N$ , such that  $\Delta\lambda = 1$ . At each  $C_\alpha$ -position, we construct the local Frenet basis from the interpolated space curve,

$$\mathbf{t}(\lambda) = \frac{\dot{\mathbf{r}}(\lambda)}{|\dot{\mathbf{r}}(\lambda)|}, \quad (3)$$

$$\mathbf{n}(\lambda) = \frac{\dot{\mathbf{t}}(\lambda)}{|\dot{\mathbf{t}}(\lambda)|}, \quad (4)$$

$$\mathbf{b}(\lambda) = \mathbf{t}(\lambda) \wedge \mathbf{n}(\lambda), \quad (5)$$

where  $\{\mathbf{t}, \mathbf{n}, \mathbf{b}\}$  are, respectively, the tangent vector, the normal vector, and the bi-normal vector to the curve. The dot denotes a derivative with respect to  $\lambda$ . Interpolating the space curve around each  $C_\alpha$ -position with a second order polynomial involving the respective left and right neighbors, we obtain

$$\dot{\mathbf{r}}(\lambda_j) = \frac{\mathbf{R}_{j+1} - \mathbf{R}_{j-1}}{2\Delta\lambda}, \quad (6)$$

$$\ddot{\mathbf{r}}(\lambda_j) = \frac{\mathbf{R}_{j+1} - 2\mathbf{R}_j + \mathbf{R}_{j-1}}{\Delta\lambda^2}, \quad (7)$$

for  $j = 2, \dots, N-1$ . At the end points of the chain one can only use forward and backward differences, respectively, and a second-order interpolation of the  $C_\alpha$ -space would lead to identical  $\{\mathbf{t}, \mathbf{n}\}$ -planes at the first and last two  $C_\alpha$ -positions, which is not compatible with a helical curve. In this case we resort to third-order interpolation, such that

$$\dot{\mathbf{r}}(\lambda_1) = \frac{-11\mathbf{R}_1 + 18\mathbf{R}_2 - 9\mathbf{R}_3 + 2\mathbf{R}_4}{6\Delta\lambda}, \quad (8)$$

$$\ddot{\mathbf{r}}(\lambda_1) = \frac{2\mathbf{R}_1 - 5\mathbf{R}_2 + 4\mathbf{R}_3 - \mathbf{R}_4}{\Delta\lambda^2}, \quad (9)$$

$$\dot{\mathbf{r}}(\lambda_N) = \frac{-2\mathbf{R}_{N-3} + 9\mathbf{R}_{N-2} - 18\mathbf{R}_{N-1} + 11\mathbf{R}_N}{6\Delta\lambda}, \quad (10)$$

$$\ddot{\mathbf{r}}(\lambda_N) = \frac{-\mathbf{R}_{N-3} + 4\mathbf{R}_{N-2} - 5\mathbf{R}_{N-1} + 2\mathbf{R}_N}{\Delta\lambda^2}. \quad (11)$$

We note here that the Frenet frames constructed at the  $C_\alpha$ -positions 2– $N$  are identical with the so-called “discrete Frenet Frames” introduced in Ref. [16].

### B. Relating Frenet frames by screw motions

Having constructed the Frenet frames, the next step consists in constructing the screw motions which link consecutive frames along the protein main chain. For this purpose, the basis vectors  $\{\mathbf{t}(\lambda_j), \mathbf{n}(\lambda_j), \mathbf{b}(\lambda_j)\} \equiv \{\mathbf{t}_j, \mathbf{n}_j, \mathbf{b}_j\}$  must be referred to their respective anchor points,  $\mathbf{R}_j$ . Defining

$$\boldsymbol{\epsilon}_j^{(1)} = \mathbf{t}_j, \quad \boldsymbol{\epsilon}_j^{(2)} = \mathbf{n}_j, \quad \boldsymbol{\epsilon}_j^{(3)} = \mathbf{b}_j, \quad (12)$$

the “tips” of the Frenet basis vectors are located at

$$\mathbf{x}_j^{(k)} = \mathbf{R}_j + \boldsymbol{\epsilon}_j^{(k)} \quad (k = 1, 2, 3), \quad (13)$$

and the mathematical problem consists in finding the screw parameters for the mappings  $\{\mathbf{x}_j^{(k)}\} \rightarrow \{\mathbf{x}_{j+1}^{(k)}\}$  for  $j = 1, \dots, N-1$ .

#### 1. Screw motions

In general, a rigid body displacement  $\mathbf{x} \rightarrow \mathbf{y}$  can be expressed in the form

$$\mathbf{y} = \mathbf{x}^{(c)} + \mathbf{D} \cdot (\mathbf{x} - \mathbf{x}^{(c)}) + \mathbf{t}, \quad (14)$$

where  $\mathbf{x}^{(c)}$  is the center of rotation,  $\mathbf{D}$  is a rotation matrix, and  $\mathbf{t}$  a translation vector. By construction,

$$\mathbf{t} = \mathbf{y}^{(c)} - \mathbf{x}^{(c)}. \quad (15)$$

The elements of the rotation matrix can be expressed in terms of three independent real parameters. One possible choice is to use the rotation angle,  $\phi$ , and the unit vector,  $\mathbf{n}$ , pointing into the direction of the rotation axis. For this parametrization,  $\mathbf{D}$  has the form[17]

$$\mathbf{D}(\mathbf{n}, \phi) = \cos \phi \mathbf{1} + (1 - \cos \phi) \mathbf{P} + \sin \phi \mathbf{N}, \quad (16)$$

where  $\mathbf{P} = (n_i n_j)$  ( $i, j = 1, 2, 3$ ) is the projector on  $\mathbf{n}$  and  $\mathbf{N}$  is a skew-symmetric  $3 \times 3$  matrix which is defined by the relation  $\mathbf{N} \cdot \mathbf{v} = \mathbf{n} \wedge \mathbf{v}$  for an arbitrary vector  $\mathbf{v}$ . The elements of  $\mathbf{N}$  are  $N_{ij} = -\sum_k \epsilon_{ijk} n_k$ , where  $\epsilon_{ijk}$  ( $i, j, k = 1, 2, 3$ ) are the components of the totally antisymmetric Levi-Civita tensor. We recall that  $\epsilon_{ijk} = \pm 1$  for, respectively, an even and odd permutation of 123, and  $\epsilon_{ijk} = 0$  zero otherwise. The parameters of the rigid-body displacement (14) depend on the choice of the rotation center,  $\mathbf{x}^{(c)}$ , and there is a special choice,  $\mathbf{x}^{(c)} = \mathbf{s}$ , for which the translation vector  $\mathbf{t}$  points into the direction of the rotation axis  $\mathbf{n}$ , such that  $\mathbf{t} \cdot \mathbf{n} > 0$ . This is known as Chasles’ theorem [18] and

the corresponding rigid body displacement describes a screw motion,

$$\mathbf{y} = \mathbf{s} + \mathbf{D}(\mathbf{n}, \phi) \cdot (\mathbf{x} - \mathbf{s}) + \alpha \mathbf{n}. \quad (17)$$

Using that  $\mathbf{D}(\mathbf{n}, \phi) \cdot \mathbf{n} = \mathbf{n}$ , one shows easily that  $\alpha$  is the projection of the translation vector on the rotation axis,

$$\alpha = \mathbf{t} \cdot \mathbf{n}. \quad (18)$$

The position  $\mathbf{s}$  is not uniquely defined, but stands for all points on the screw axis. Defining  $\mathbf{s}^{(c)}$  to be the point for which the distance  $|\mathbf{s} - \mathbf{x}^{(c)}|$  is a minimum, the screw axis is defined through

$$\mathbf{s} = \mathbf{s}^{(c)} + \mu \mathbf{n}, \quad -\infty < \mu < +\infty, \quad (19)$$

where

$$\mathbf{s}^{(c)} = \mathbf{x}^{(c)} + \frac{1}{2}(\mathbf{t}^\perp + \cos(\phi/2)\mathbf{n} \wedge \mathbf{t}), \quad (20)$$

and  $\mathbf{t}^\perp = \mathbf{t} - (\mathbf{n} \cdot \mathbf{t})\mathbf{n}$  is the component of  $\mathbf{t}$  which is perpendicular to the rotation axis. We note that  $(\mathbf{s}^{(c)} - \mathbf{x}^{(c)}) \cdot \mathbf{n} = 0$ . The radius of the screw motion is defined through  $\rho = |\mathbf{x}^{(c)} - \mathbf{s}^{(c)}|$  and it follows from (20) that

$$\rho = \frac{|\mathbf{t}^\perp|}{2} \sqrt{1 + \cot(\phi/2)^2}. \quad (21)$$

## 2. Determining the screw parameters

Assuming that the Frenet frames at the  $C_\alpha$ -positions have been constructed, the fold of a protein is defined by the sequence of screw motions  $\mathbf{x}_j^{(k)} \rightarrow \mathbf{x}_{j+1}^{(k)}$ , where

$$\mathbf{x}_{j+1}^{(k)} = \mathbf{s}_j^{(c)} + \mathbf{D}(\mathbf{n}_j, \phi_j) \cdot (\mathbf{x}_j^{(k)} - \mathbf{s}_j^{(c)}) + \alpha_j \mathbf{n}_j, \quad (22)$$

for  $j = 1, \dots, n-1$  and  $k = 1, 2, 3$ . The corresponding parameters are computed as follows:

1. Determine the translation vectors

$$\mathbf{t}_j = \mathbf{R}_{j+1} - \mathbf{R}_j. \quad (23)$$

2. Perform a rotational least squares fit[19]  $\{\epsilon_j^{(k)}\} \rightarrow \{\epsilon_{j+1}^{(k)}\}$  by minimizing the target function

$$m(Q_j) = \sum_{k=1}^3 \left| \epsilon_{j+1}^{(k)} - \mathbf{D}(Q_j) \cdot \epsilon_j^{(k)} \right|^2 \quad (24)$$

with respect to four quaternion parameters,  $Q = \{q_0, q_1, q_2, q_3\}$ , which parametrize the rotation matrix according to

$$\mathbf{D}(Q) = \begin{pmatrix} q_0^2 + q_1^2 - q_2^2 - q_3^2 & 2(q_1 q_2 - q_0 q_3) & 2(q_0 q_2 + q_1 q_3) \\ 2(q_1 q_2 + q_0 q_3) & q_0^2 - q_1^2 + q_2^2 - q_3^2 & -2(q_0 q_1 - q_2 q_3) \\ -2(q_0 q_2 - q_1 q_3) & 2(q_0 q_1 + q_2 q_3) & q_0^2 - q_1^2 - q_2^2 + q_3^2 \end{pmatrix}. \quad (25)$$

The quaternion parameters are normalized such that  $q_0^2 + q_1^2 + q_2^2 + q_3^2 = 1$ , which leaves three free parameters describing the rotation. We note here only that the minimization of (24) leads to an eigenvector problem for the optimal quaternion, which can be efficiently solved by standard linear algebra routines, and that the corresponding eigenvalue is the squared superposition error [19]. The latter is zero for superposition of Frenet frames, since two orthonormal and equally oriented vector sets can be perfectly superposed. It is also worthwhile noting that the upper limit in the sum in (24) can be changed from 3 to 2, since two linearly independent vectors with the same origin, here  $\mathbf{t}_j$  and  $\mathbf{n}_j$ , suffice to define a rigid body.

3. Extract  $\mathbf{n}_j$  and  $\phi_j$  from the quaternion parameters  $Q_j$ . This can be easily achieved by exploiting the relations

$$\left. \begin{aligned} q_0 &= \cos(\phi/2) \\ q_1 &= \sin(\phi/2)n_x \\ q_2 &= \sin(\phi/2)n_y \\ q_3 &= \sin(\phi/2)n_z \end{aligned} \right\} \quad (26)$$

Here and in the following the index  $j$  is dropped. Several cases have to be considered. If  $\sqrt{q_1^2 + q_2^2 + q_3^2} > \epsilon$ , where  $\epsilon$  depends on the machine precision of the computer being used, we compute a “tentative rotation axis”

$$\mathbf{n}_t = \frac{1}{\sqrt{q_1^2 + q_2^2 + q_3^2}} \begin{pmatrix} q_1 \\ q_2 \\ q_3 \end{pmatrix}. \quad (27)$$

Then we check if  $\mathbf{t} \cdot \mathbf{n}_t \geq 0$ . If this is the case we set

$$\mathbf{n} = \mathbf{n}_t, \quad (28)$$

$$\phi = 2 \arccos(q_0). \quad (29)$$

In case that  $\mathbf{t} \cdot \mathbf{n}_t < 0$  we set

$$\mathbf{n} = -\mathbf{n}_t, \quad (30)$$

$$\phi = 2 \arccos(-q_0). \quad (31)$$

This corresponds to replacing  $Q \rightarrow -Q$  before evaluating  $\mathbf{n}$  and  $\phi$  according to (28) and (29). Such a replacement is possible since the elements of  $\mathbf{D}(Q)$  are homogeneous functions of order two in the quaternion parameters, such that  $\mathbf{D}(Q) = \mathbf{D}(-Q)$ .

For the sake of completeness, we finally mention the case that  $\sqrt{q_1^2 + q_2^2 + q_3^2} \leq \epsilon$ , which corresponds to a pure translation and cannot occur in our application to protein backbones. In this case one would set  $\phi = 0$  and  $\mathbf{n} = \mathbf{t}/|\mathbf{t}|$ .

4. Using the parameters  $\{\mathbf{n}_j, \phi_j\}$  and defining the positions  $\mathbf{R}_j$  to be the rotation centers,  $\mathbf{x}^{(c)} = \mathbf{R}_j$ , compute for  $j = 1, \dots, N-1$

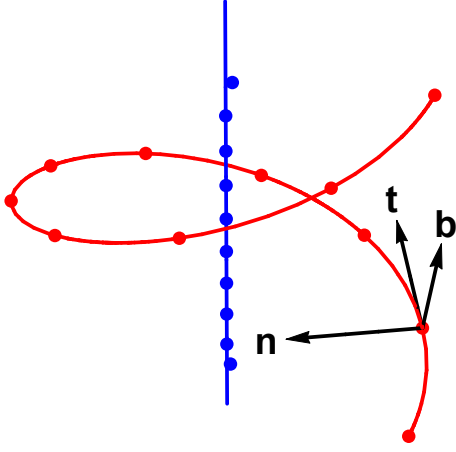


FIG. 1: Frenet frame  $\{t, n, b\}$  at one point of the helicoidal curve defined in Eq. (33) (red solid line). Setting  $R = 1$  and  $h = 0.3$ , the latter is shown for one turn, together with  $N = 11$  equidistantly spaced sampling points (red points). The blue line is the helix axis and the blue points correspond to the rotation centers  $s_j^{(c)}$  ( $j = 1, \dots, N - 1$ ). The figure has been produced with the Mathematica software [20].

- (a) the positions  $s_j^{(c)}$  on the local screw axes according to relation (20),
- (b) the local helix radii according to relation (21).

### 3. Regularity of PSSEs

To quantify the regularity of PSSEs, we introduce the distance measure

$$\delta(j) = \left| s_j^{(c)} + t_j^{\parallel} - s_{j+1}^{(c)} \right|, \quad j = 1, \dots, N - 2, \quad (32)$$

where  $t_j^{\parallel} = \mathbf{n} \cdot \mathbf{t}_j$ . For an ideal PSSE, where all consecutive Frenet frames are related by the same screw motion,  $\delta(j)$  is strictly zero. This measure of non-ideality deviates from the “straightness” parameter in the ScrewFit algorithm [5], which is defined as  $\sigma_j = \mu_{j+1} \cdot \mu_j$  with  $\mu_j = s_{j+1}^{(c)} - s_j^{(c)}$ , and which defines ideality of PSSEs through the cosine of the angle between subsequent local screw axes.

### C. Numerical test

To test the numerical construction of Frenet frames, we consider a perfect helicoidal curve and compare the exact Frenet frames with the corresponding numerical approximations. The parametric representation of the curve is

$$\mathbf{r}(\lambda) = \rho \cos(\lambda) \mathbf{e}^{(x)} + \rho \sin(\lambda) \mathbf{e}^{(y)} + h\lambda \mathbf{e}^{(z)}, \quad (33)$$

where  $\rho > 0$  is the radius of the helix and its pitch is  $p = h/2\pi$ . Fig. 1 shows the form of the curve (33) for

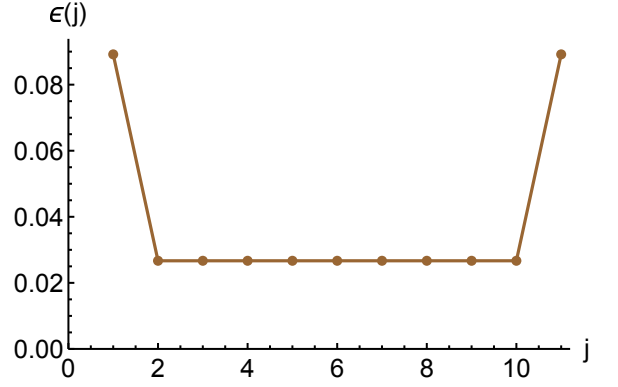


FIG. 2: Overlap error (35) for the bases  $\tilde{\mathbf{F}}(\lambda_j)$  and  $\mathbf{F}(\lambda_j)$  at the red points in Fig. 1.

one complete turn (red line), setting  $R = 1$  and  $h = 0.3$  in arbitrary length units. Defining the matrix  $\mathbf{F}(\lambda) = (\mathbf{t}(\lambda), \mathbf{n}(\lambda), \mathbf{b}(\lambda))$ , it follows from (33) that

$$\mathbf{F}(\lambda) = \begin{pmatrix} -\frac{R \sin(\lambda)}{\sqrt{h^2 + R^2}} & -\cos(\lambda) & \frac{h \sin(\lambda)}{\sqrt{h^2 + R^2}} \\ \frac{R \cos(\lambda)}{\sqrt{h^2 + R^2}} & -\sin(\lambda) & -\frac{h \cos(\lambda)}{\sqrt{h^2 + R^2}} \\ \frac{h}{\sqrt{h^2 + R^2}} & 0 & \frac{R}{\sqrt{h^2 + R^2}} \end{pmatrix}. \quad (34)$$

Using the method described in Section II A, we construct numerical approximations  $\tilde{\mathbf{F}}(\lambda_j)$  of the Frenet bases (34) at  $N = 11$  equidistant sampling points,  $\mathbf{R}_j$ , which are shown as red dots in Fig. 1. From these Frenet bases we construct the axis points  $s_j^{(c)}$  (blue dots), which are shown together with the exact screw axis (blue line). For the first and the last axis point one notices a visible offset from the latter. We quantify the error of the numerically computed Frenet bases,  $\tilde{\mathbf{F}}(\lambda_j)$ , as

$$\epsilon(j) = \sqrt{\text{tr} \{ \Delta(j)^T \cdot \Delta(j) \}}, \quad (35)$$

where

$$\Delta(j) = \tilde{\mathbf{F}}(\lambda_j)^T \cdot \mathbf{F}(\lambda_j) - \mathbf{1}. \quad (36)$$

For a perfect overlap of  $\tilde{\mathbf{F}}(\lambda_j)$  and  $\mathbf{F}(\lambda_j)$  one should have  $\tilde{\mathbf{F}}(\lambda_j)^T \cdot \mathbf{F}(\lambda_j) = \mathbf{1}$ , such that  $\epsilon(j) = 0$ . We note that  $\epsilon(j)$  is the Frobenius norm [21] of  $\Delta(j)$ . Fig. 2 shows  $\epsilon(j)$  corresponding to the Frenet basis in Fig. 1 and confirms the slight offset of the first and last axis point from the ideal screw axis.

## III. APPLICATIONS

In the following we consider two applications of the coarse-grained model for protein secondary structure, which has been described in the previous section and which will be referred to as ScrewFrame in the following. The first application concerns the construction of a



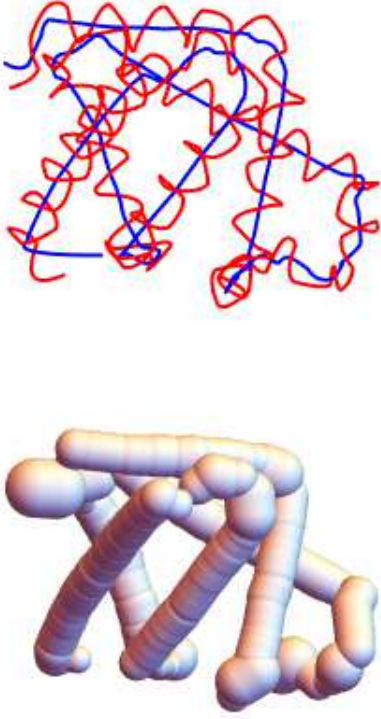


FIG. 3: **Top:**  $C_\alpha$ -curve (red) of myoglobin (PDB code 1A6G) and B-spline curve (blue) linking the screw motion centers  $\{s_j^{(c)}\}$ . **Bottom:** Tube representation of the  $C_\alpha$ -curve. The local tube radii equal the respective helix radii  $\{\rho_j\}$  of the screw motions linking the Frenet frames  $j$  and  $j + 1$  ( $j = 1, \dots, N - 1$ ). The figure has been produced with the Mathematica software [20].

tube model for a protein from the ScrewFrame parameters and in the second application, these parameters are used for a comparative study of ScrewFrame and DSSP for secondary structure assignment.

#### A. Tube representation of a protein

As a first application we consider the ScrewFrame model for myoglobin, which is an oxygen-binding protein in muscular tissues. Myoglobin is composed of 151 amino acids which fold into a globular form and the dominant PSSEs are  $\alpha$ -helices. For our demonstration we use the crystallographic structure 1A6G of the Protein Data Bank [22]. The red and blue line in the upper part of Fig. 3 display, respectively, the space curve defined by the positions  $\mathbf{R}_j$  of the  $C_\alpha$ -atoms and the space curve linking the corresponding screw motion centers  $s_j^{(c)}$ . Both space curves are constructed by a piecewise polynomial interpolation of second order [20]. The blue line indicates the global fold of the protein, where ideal PSSEs appear simply as straight segments. In the following we refer to this line as the protein screw axis. It plays the same role as the “overall protein axis” in

	$\alpha$ -helix	$\beta$ -strand $\uparrow\uparrow$	$\beta$ -strand $\uparrow\downarrow$	3-10 helix	$\pi$ -helix
ScrewFit	0.165	0.061	0.051	0.122	0.165
ScrewFrame	0.227	0.098	0.080	0.187	0.227

TABLE I: Screw radii in nm for standard model structures generated with Chimera [23]. Since ScrewFit uses the  $C$ -atoms in the peptide planes as reference points for the (pure) rotations, whereas as ScrewFrame uses the  $C_\alpha$ -atoms, the radii determined by ScrewFit are systematically smaller than those obtained from ScrewFrame.

the P-Curve algorithm [8], although its construction is different. The lower part of the figure shows the corresponding “tube model”, where the axis of the tube equals the protein screw axis and the local tube radius corresponds to the radius of the local screw motion. As in the original ScrewFit algorithm, the screw radius allows for a discrimination of different types of PSSEs (see Table I). Fig. 4 displays this quantity for myoglobin as a function of the residue number (blue line) and, for comparison, the corresponding values for the ScrewFit algorithm (brown line). The light gray stripes indicate  $\alpha$ -helices found by the DSSP algorithm. The comparison of the results with the ScrewFit analysis of the same protein structure shows that both methods indicate  $\alpha$ -helices in the same place, in close agreement with DSSP. Here it must be observed that the definition of the screw radii is not the same for ScrewFit and ScrewFrame. The rotation centers in the ScrewFit algorithm are the  $C$ -atoms in the  $C - O - N$ -peptide planes, whereas the  $C_\alpha$ -atoms are used for ScrewFrame. For an ideal  $\alpha$ -helix the corresponding radii are 0.165 nm and 0.227 nm, respectively (see Table I). Fig. 5 shows the regularity measure (32) which plays an important role in the attribution of secondary structure elements to be discussed in the following section.

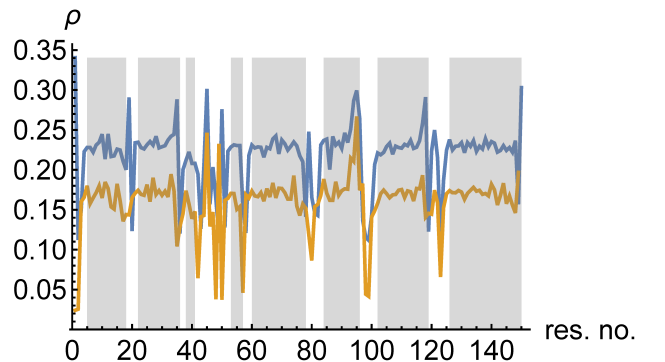


FIG. 4: The radius  $\rho$  for the ScrewFrame representation (blue line) of myoglobin (PDB code 1A6G) as a function of the residue number and the corresponding values for ScrewFit (brown line). The light gray stripes indicate the  $\alpha$ -helices found by DSSP.

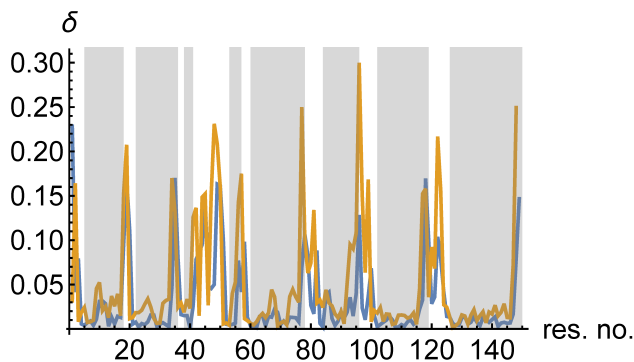


FIG. 5: The regularity measure (32) for the ScrewFrame representation of myoglobin (PDB code 1A6G) as a function of the residue number and the corresponding values for ScrewFit (brown line). The light gray stripes indicate the  $\alpha$ -helices found by DSSP.

### B. Analysis of the ASTRAL database

In order to compare our  $C_\alpha$  based helicoidal analysis with the original ScrewFit method based on peptide planes [5, 6], we applied both methods to the “all  $\alpha$ ” and “all  $\beta$ ” categories of the ASTRAL subset of the SCOPE database [24], using the ASTRAL SCOPE 2.04 subset with less than 40% sequence identity. In order to be able to work efficiently with such a large collection of protein structures, we constructed an ActivePaper [25] containing the structures of the ASTRAL entries in MOSAIC format [26]. This file is available for download [27]. In addition to the ASTRAL database of real protein structures, we use ideal secondary-structure elements ( $\alpha$ -helix,  $\pi$ -helix, 3 – 10-helix, parallel and anti-parallel  $\beta$ -strands) for polyalanine, which were constructed using the program Chimera [23].

We also compare to DSSP secondary structure assignments for this database, using our own implementation of the DSSP algorithm which follows the description in the original publication [3] but, like the current version 2 of the DSSP software [28], computes an ideal position for the backbone hydrogen positions instead of using experimental values, even if the latter are available.

As a first step, we compute ScrewFit and ScrewFrame parameters for all structures in the all- $\alpha$  and all- $\beta$  subsets of the ASTRAL database. In order to avoid inaccuracies introduced by the third-order approximations given by Eqs. (8)–(11), we do not compute Frenet frames for the first and last residue of each chain. For structures with missing residues, we compute the parameters for each continuous chain segment separately. Since the input structures are dominated by  $\alpha$ -helices and  $\beta$ -strands, respectively, we expect the distribution of our parameters to show clear peaks that correspond to these secondary structure elements.

The most important helix parameter for secondary structure description is the helix radius  $\rho$ , whose dis-

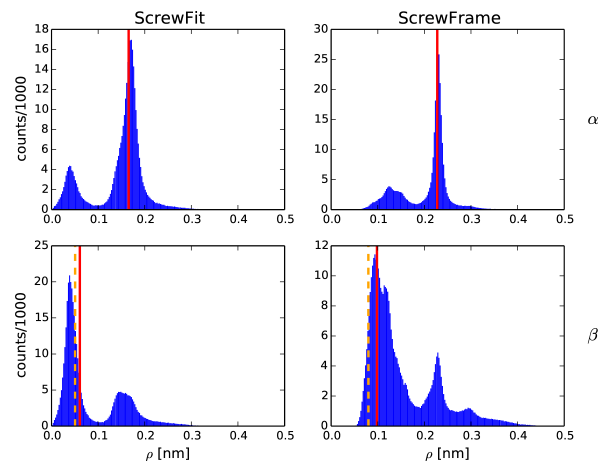


FIG. 6: The helix radius  $\rho$  for the all- $\alpha$  (top) and the all- $\beta$  structures (bottom), using the ScrewFit (left) and ScrewFrame (right) methods. Note that the ScrewFit radius is based on the C-atoms, whereas the ScrewFrame radius corresponds to the  $C_\alpha$ -atoms, which explains the different values. The vertical lines indicate the values for ideal secondary-structure elements. For  $\beta$ -strands, there are two ideal values, one for parallel (red, drawn-out) and one for antiparallel (orange, dashed) strands.

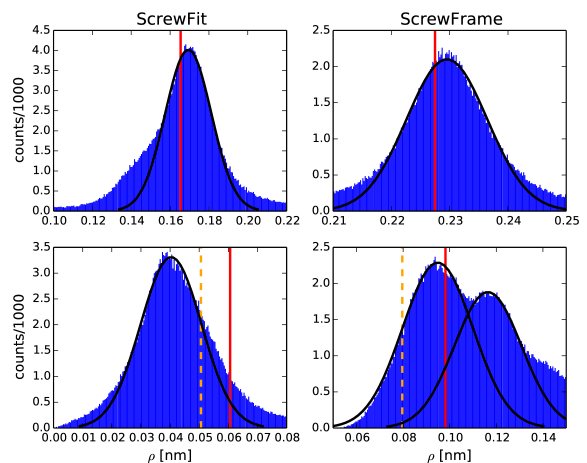


FIG. 7: The helix radius  $\rho$  around the ideal- $\alpha$  value for the all- $\alpha$  subset (top) and around the ideal- $\beta$  values for the all- $\beta$  structures (bottom), using the ScrewFit (left) and ScrewFrame (right) methods. The vertical lines indicate values for ideal secondary-structure elements, as in Fig. 6. The Gaussian distributions fitted to the peaks are drawn in black, their parameters are given in Table II. The  $\beta$  distribution for ScrewFrame can be well described as a superposition of two Gaussian distributions, corresponding to parallel and antiparallel strands. The ScrewFit method cannot resolve this difference.

tribution in the ASTRAL database is shown in Fig. 6. The vertical lines show for comparison the values for ideal  $\alpha$ -helices and  $\beta$ -strands. For the  $\beta$ -strands, the red drawn-out lines stand for parallel and the orange dashed lines for antiparallel strands. A more detailed

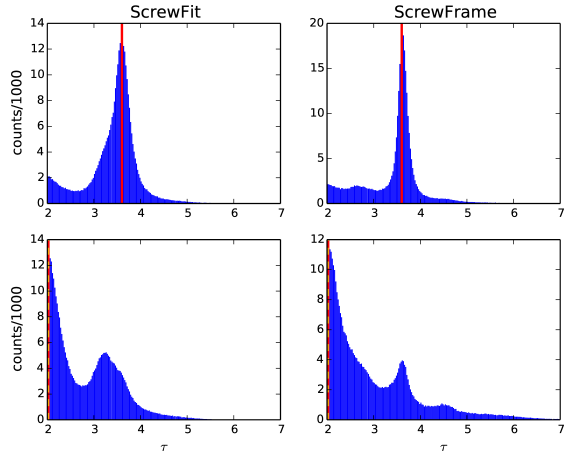


FIG. 8: The number of amino acid residues per full turn,  $\tau$ , for the all- $\alpha$  (top) and the all- $\beta$  structures (bottom) using the ScrewFit (left) and ScrewFrame (right) methods. The theoretical minimal value of  $\tau = 2$  is very close to the observed value for  $\beta$ -sheets.

view is given in Fig. 7, which shows only the region around the dominant peak for each histogram, together with Gaussian distributions fitted to the peaks. The peaks are rather well described by a Gaussian, and the ScrewFrame method even allows to resolve the difference between parallel and antiparallel  $\beta$ -strands.

Whereas the average  $\rho$  value for  $\alpha$ -helices is close to the value for an ideal helix, this is not the case at all for  $\beta$ -strands. This can be understood by looking at the distribution of the number of amino acids per full turn,  $\tau$ , shown in Fig. 8. Since the rotation angle is by definition in the interval  $[-\pi \dots \pi]$ , the minimal value of  $\tau$  is 2. This is also the value that describes an ideal  $\beta$ -strand, which is a flat structure. Any deviation from the ideal  $\beta$ -strand has a larger  $\tau$ , and because  $\rho$  and  $\tau$  are not independent (the length of the curve arc linking two neighboring  $C_\alpha$  atoms is nearly constant), the deviation in  $\rho$  from the ideal value is asymmetric as well.

The regularity measure  $\delta$ , defined in Eq. (32), is shown in Fig. 9. It shows that the ScrewFrame secondary structure elements are more regular than those identified by ScrewFit, in particular for structures dominated by  $\alpha$ -helices. We do not show here the distributions of the other parameters defined in the initial ScrewFit publication [5], but they are included in the electronic supplementary material. We note that the parameter distributions are in general narrower and thus better defined for ScrewFrame than for ScrewFit. We attribute this fact to fluctuations in the orientations of the peptide plans that have no impact on the  $C_\alpha$  geometry.

We use the Gaussian distributions shown in Fig. 7 as the basis for defining secondary-structure elements. We define an  $\alpha$ -helix as a sequence of at least four consecu-

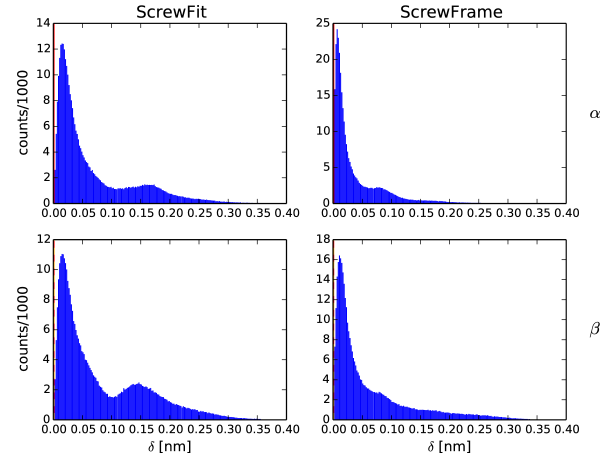


FIG. 9: The regularity measure  $\delta$  defined in Eq. (32) for the all- $\alpha$  and the all- $\beta$  subset of the ASTRAL data base (top and bottom, respectively).

tive  $C_\alpha$  atoms whose screw transformations satisfy

$$\frac{|\rho - \mu_\rho|}{\sigma_\rho} < 3 \quad (37)$$

$$\delta < 0.02 \text{ nm} \quad (38)$$

where  $\mu_\rho$  and  $\sigma_\rho$  are the mean value and standard deviation of the Gaussian distribution for the  $\alpha$  peak in Fig. 7. The numerical values of these parameters are shown in Table II.

	$\alpha$ -helix	$\beta$ -strand $\uparrow\uparrow$	$\beta$ -strand $\uparrow\downarrow$
$\mu_\rho$	0.230	0.116	0.095
$\sigma_\rho$	0.007	0.014	0.015

TABLE II: The parameters of the Gaussians fitted to the peaks in the distributions of the ScrewFrame parameter  $\rho$  (see Fig. 7). All values are in units of nm.

We define a  $\beta$ -strand as a segment of consecutive  $C_\alpha$  atoms whose screw transformations satisfy

$$\min \left( \frac{|\rho - \mu_\rho^{(1)}|}{\sigma_\rho^{(1)}}, \frac{|\rho - \mu_\rho^{(2)}|}{\sigma_\rho^{(2)}} \right) < 1 \quad (39)$$

$$\delta < 0.08 \text{ nm} \quad (40)$$

where  $\mu_\rho^{(1/2)}$  and  $\sigma_\rho^{(1/2)}$  are the mean values and standard deviations of the Gaussian distributions for the parallel and antiparallel  $\beta$  peaks in Fig. 7. The numerical parameters in these definitions were chosen to make our definitions match the secondary structure assignments made by the DSSP method.

There is a fundamental difference between our approach and the DSSP method for defining  $\beta$ -strands. The ScrewFrame approach looks for a regular structure along the peptide chain, whereas the DSSP method

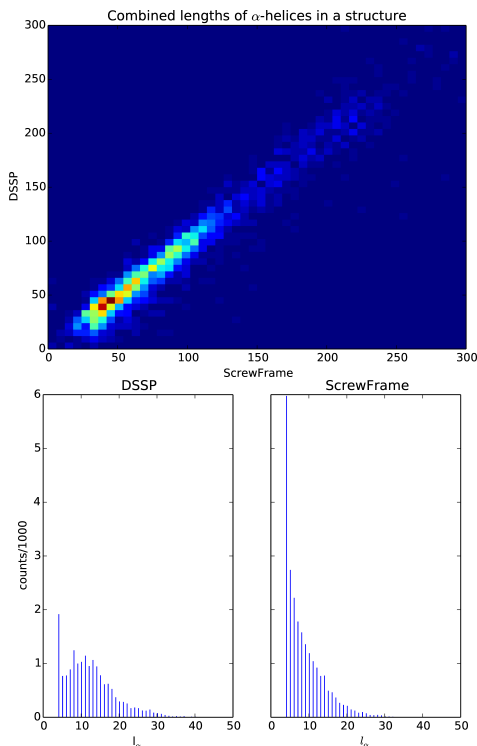


FIG. 10: **Top:** A two-dimensional histogram comparing the total number of residues inside  $\alpha$ -helices as identified by ScrewFrame and DSSP. The strong localization of the distribution around the diagonal shows the similarity between these two assignments. **Bottom:** The distribution of the lengths of identified  $\alpha$ -helices, left for DSSP, right for ScrewFrame. The fatter tail for DSSP and the larger number of short helices for ScrewFrame are due to the fact that ScrewFrame breaks up strongly deformed helices into several pieces, whereas DSSP considers them a single helix.

identifies hydrogen bonds between the strands that make up a  $\beta$ -sheet. ScrewFrame thus finds individual strands, which can be paired up to identify sheets in a separate step. A strand must consist of at least three consecutive residues in order to be considered regular; in fact, the regularity measure  $\delta$  is defined in terms of the difference of two consecutive screw transformations, each of which connects two residues. DSSP needs to look at two strands simultaneously in order to identify  $\beta$  structures, but has no minimal length condition and in fact admits  $\beta$ -sheets as small as a single h-bonded residue pair. For practically relevant  $\beta$ -sheets in real protein structures, these differences are, however, not important, but they must be understood for interpreting the following comparison between the two methods.

A one-to-one comparison of secondary structure elements from two different assignment methods is not of particular interest, because an exact match is the exception rather than the rule. The inherent fuzziness of secondary structure definitions leads to arbitrary choices and thus inevitable differences. The most frequent de-

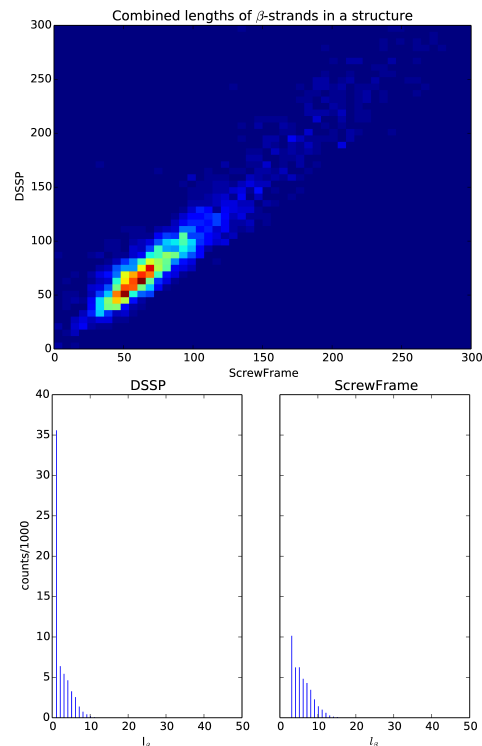


FIG. 11: **Top:** A two-dimensional histogram comparing the total number of residues inside  $\beta$ -strands as identified by ScrewFrame and DSSP. The strong localization of the distribution around the diagonal shows the similarity between these two assignments. **Bottom:** The distribution of the lengths of identified  $\beta$ -helices, left for DSSP, right for ScrewFrame. The peak at very short strands in the DSSP distribution is absent from the ScrewFrame results because ScrewFrame needs at least three consecutive residues to recognize a regular structure.

viation between two assignments is the end points of secondary structure elements, where a difference of one or two residues is common and acceptable. Another frequent deviation concerns deformed secondary structure elements, which one method may identify as a single element whereas another one recognizes it as multiple distinct elements.

We therefore chose a statistical comparison to compare the ScrewFrame results to those of DSSP, which is shown in Figs. 10 for  $\alpha$ -helices and 11 for  $\beta$ -strands. We consider two quantities: (1) the total number of residues of a given structure which are inside a recognized secondary-structure element, and (2) the length of each individual secondary-structure element. We compute the first quantity for both methods and show their joint distribution (upper plot in the two figures). For the vast majority of structures, the two residue counts are close to equal, which means that neither method yields systematically more or longer secondary-structure elements than the other. The lower plots show the distributions of the lengths of individual secondary-structure



elements. For  $\alpha$ -helices, DSSP has a fatter tail (helices of length 20 or more), whereas ScrewFrame identifies a larger number of short helices. The reason for these differences is that ScrewFrame tends to split up kinked helices which DSSP identifies as single units. For  $\beta$ -strands, we notice that DSSP identifies many more very short elements. This is due to the different definitions: a single  $\beta$ -type hydrogen bond is sufficient to define a  $\beta$ -sheet in DSSP, but ScrewFrame requires at least three consecutive residues to identify any regular structure.

#### IV. CONCLUSION AND OUTLOOK

We have presented a generalization of the ScrewFit method for protein structure assignment and description, which uses only the positions of the  $C_\alpha$ -atoms along the protein backbone. As in the ScrewFit approach, the global protein fold is described as a succession of screw motions relating consecutive recurrent motifs along the protein backbone, but the “motifs” are here the tripods (planes) formed by the three (two) orthonormal vectors of the local Frenet bases to the  $C_\alpha$  space curve. Despite the fact that ScrewFrame uses less information than ScrewFit, all standard PSSEs are recognized on the basis of thresholds for the local screw radii and a suitably defined regularity measure. ScrewFrame even permits to distinguish between parallel and antiparallel  $\beta$ -strands, which the classical

ScrewFit method fails to do. A thorough comparison with the commonly used DSSP method on the assignment of PSSEs in the ASTRAL database shows that both methods yield very similar results for the total amount of PSSEs. ScrewFrame tends, however, to break long helices into smaller pieces, such that the length distribution of PSSEs is different. Due to the minimalistic character of the geometrical model for protein folds, the evaluation of the ScrewFrame model parameters is very efficient. This allows for working with protein structure databases and for analyzing simulated molecular dynamics trajectories of proteins. ScrewFrame may also be used as a starting point for the development of minimalistic models for protein structure and dynamics, similar to the wormlike chain model [29], which has been successfully applied to DNA [30]. As already mentioned, our method can also be used to analyze dynamical processes, such as the folding and unfolding of peptides [31] and it can describe the fold of intrinsically disordered proteins.

An ActivePaper [25] containing all the software, input datasets, and results from this study is available as supplementary material. The datasets can be inspected with any HDF5-compatible software, e.g. the free HDFView.[32] Running the programs on different input data requires the ActivePaper software [25].

- 
- [1] L. Pauling and R. B. Corey, *P Natl Acad Sci USA* **37**, 729 (1951).
  - [2] L. Pauling, R. B. Corey, and H. R. Branson, *P Natl Acad Sci USA* **37**, 205 (1951).
  - [3] W. Kabsch and C. Sander, *Biopolymers* **22**, 2577 (1983).
  - [4] D. Frishman and P. Argos, *Proteins* **23**, 566 (1995).
  - [5] G. R. Kneller and P. Calligari, *Acta Crystallogr D* **62**, 302 (2006).
  - [6] P. A. Calligari and G. R. Kneller, *Acta Crystallogr D* **68**, 1690 (2012).
  - [7] J.-M. Chandonia, G. Hon, N. S. Walker, L. Lo Conte, P. Koehl, M. Levitt, and S. E. Brenner, *Nucleic Acids Research* **32**, D189 (2004).
  - [8] H. Sklenar, C. Etchebest, and R. Lavery, *Proteins* **6**, 46 (1989).
  - [9] J. M. Grimes, S. D. Fuller, and D. I. Stuart, *Acta Crystallogr D* **55**, 1742 (1999).
  - [10] R. Marabini, J. R. Macias, J. Vargas, A. Quintana, C. O. S. Sorzano, and J. M. Carazo, *Acta Crystallogr D* **69**, 695 (2013).
  - [11] V. Tozzini, *Curr. Opin. Struct. Biol.* **15**, 144 (2005).
  - [12] M. Levitt and J. Greer, *J Mol Biol* **114**, 181 (1977).
  - [13] F. Dupuis, J.-F. Sadoc, and J.-P. Mornon, *Proteins* **55**, 519 (2004).
  - [14] G. Labesse, N. Colloc'h, J. Pothier, and J.-P. Mornon, *Computer applications in the biosciences: CABIOS* **13**, 291 (1997).
  - [15] S.-Y. Park, M.-J. Yoo, J.-M. Shin, and K.-H. Cho, *BMB Reports* **44**, 118 (2011).
  - [16] S. Hu, M. Lundgren, and A. J. Niemi, *Phys Rev E* **83**, 061908 (2011).
  - [17] S. Altman, *Rotations, Quaternions, and Double Groups* (Clarendon Press, Oxford, 1986).
  - [18] M. Chasles, *Bulletin des Sciences Mathématiques, Astronomiques, Physiques et Chimiques* **14**, 321 (1830).
  - [19] G. R. Kneller, *Mol Simulat* **7**, 113 (1991).
  - [20] Wolfram Research Inc., *Mathematica*, Version 10.0 (Wolfram Research Inc., Champaign, Illinois, USA, 2014).
  - [21] G. Golub and C. van Loan, *Matrix Computations* (The John Hopkins University Press, 1996).
  - [22] J. Kirchmair, P. Markt, S. Distinto, D. Schuster, G. Spitzer, K. Liedl, T. Langer, and G. Wolber, *J. Med. Chem* **51**, 7021 (2008).
  - [23] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin, *J Comput Chem* **25**, 1605 (2004).
  - [24] N. K. Fox, S. E. Brenner, and J. M. Chandonia, *Nucleic Acids Research* **42**, D304 (2013).
  - [25] K. Hinsén, *ActivePapers*, <http://www.activepapers.org/> (2014).
  - [26] K. Hinsén, *Journal of chemical information and modeling* **54**, 131 (2014).
  - [27] K. Hinsén, *ASTRAL-SCOPe subset 2.04 in ActivePapers format* (2014), URL <http://dx.doi.org/10.5281/zenodo.11086>.
  - [28] M. Hekkelman, *DSSP 2.2.1*,

- <http://swift.cmbi.ru.nl/gv/dssp/> (2013).
- [29] M. Doi and S. Edwards, *The Theory of Polymer Dynamics* (Oxford University Press, New York, 1986).
- [30] J. F. Marko and E. D. Siggia, *Macromolecules* **28**, 8759 (1995).
- [31] G. Spampinato and G. Maccari, *J Chem Theory Comput* **10**, 3885 (2014).
- [32] The HDF Group, *HDFView*, <http://www.hdfgroup.org/hdf-java-html/hdfview/> (2013).